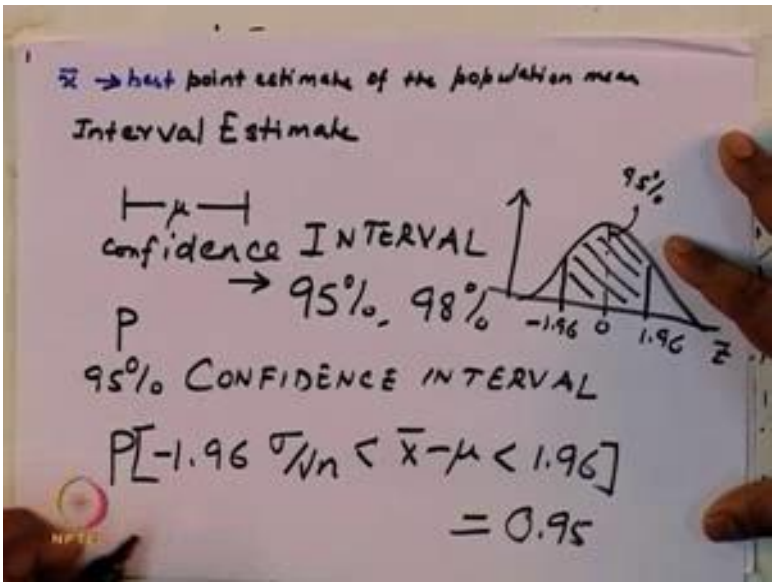
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 31
Test of Hypothesis – 1

Hello and welcome to today's class. In the last lecture we had discussed about confidence intervals, and we discussed about point estimates and interval estimates.

(Refer Slide Time: 00:36)



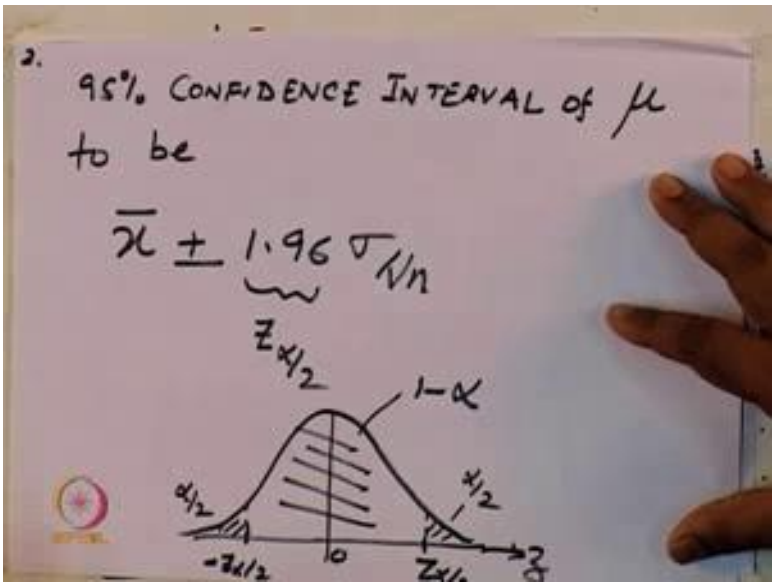
So, we had shown and we had discussed that for any sample x bar is the best point estimate of the population mean. So, what is the point estimate and interval estimate which actually prescribes a given range within which your mu is supposed to lie. So, this would be a better estimate as opposed to predicting one single value you give a range over which mule lies. And what we showed was the best way of finding out a confidence interval, so this is the confidence interval let us say this is a confidence interval. And to do this what we do is; so we can prescribe the confidence interval which is let us say 95 percent or 98 percent or any percentage for that matter.

And how do we find out the confidence interval. So, let us say if you want the confidence interval for 95 percent. So, we write down probability of minus 1.96 sigma by root n less than x bar minus mu less than 1.96. So, the integral of the curve in this for the standard normal distribution you know ok; so this is 1.96 this is standard normal

value. So, this is your 0 this is minus 1.96 this is 1.96. So, within 2 standard deviations this area is 95 percent or will return the probability of 0.95.

From this expression we can find out the 95 percent confidence interval of mu to be x bar plus minus 1.96 sigma by root n.

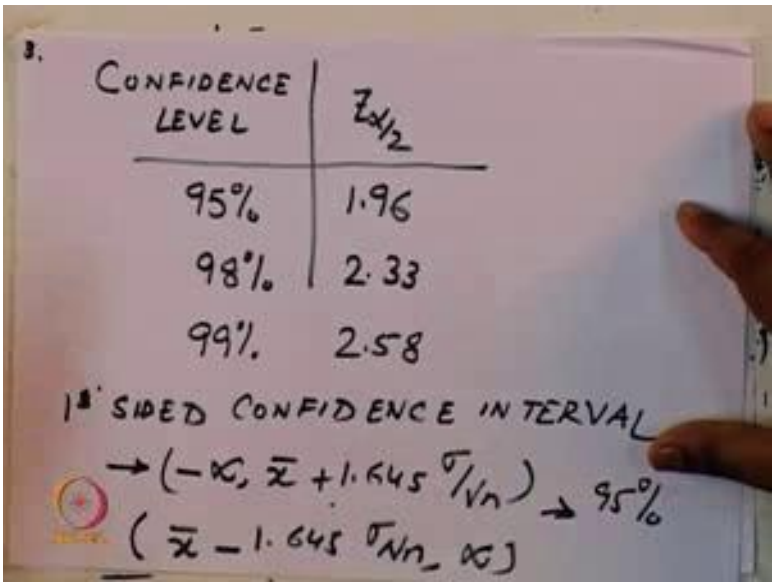
(Refer Slide Time: 03:17)



So, this term you can write it in general as Z of alpha by 2. What is Z of alpha by 2? If I plot this curve if this is my z, if this is 0 for standard normal variable, this is your standard normal variable z. This is Z of alpha by 2, this is minus Z alpha by 2, and this is plus Z alpha by 2. And if you integrate these two areas you get a value of alpha. So, this area is alpha by 2 this area is alpha by 2, so the probability under this curve will give you 1 minus of alpha.

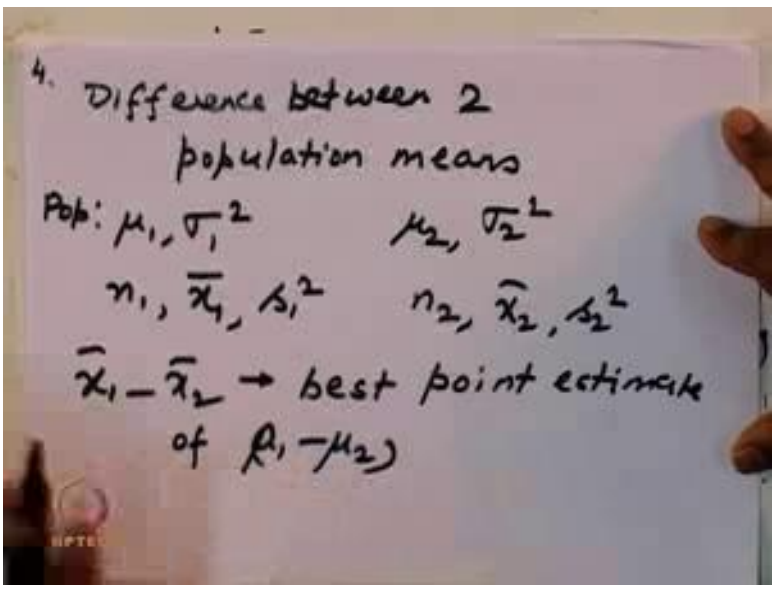
We can prescribe the value of confidence interval and get the corresponding value of Z of alpha by 2.

(Refer Slide Time: 04:38)



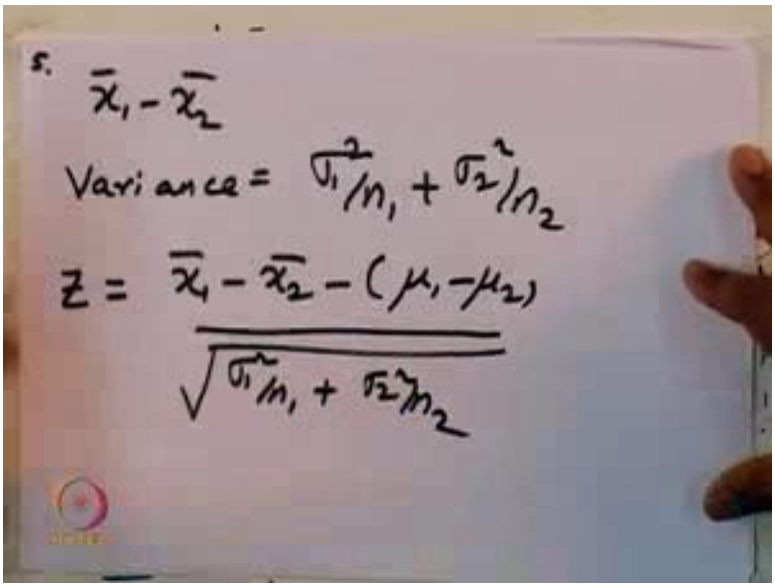
So, for 95 percent confidence interval this value is 1.96, for 98 percent this value is 2.33 and for 99 percent this value is 2.58. Now from this, this is a two sided two tail confidence interval right, you have a lower bound and an upper bound. You can also find out a minimum bound or an upper bound, a lower bound or an upper bound. In that case the confidence interval one sided constant interval is given by either minus infinity to x bar plus 1.645 sigma by root n or x bar minus 1.645 sigma by root n to plus infinity. So, this gives you the higher value, maximum value and this gives a lower value. So, this is for 95 percent confidence.

(Refer Slide Time: 06:24)



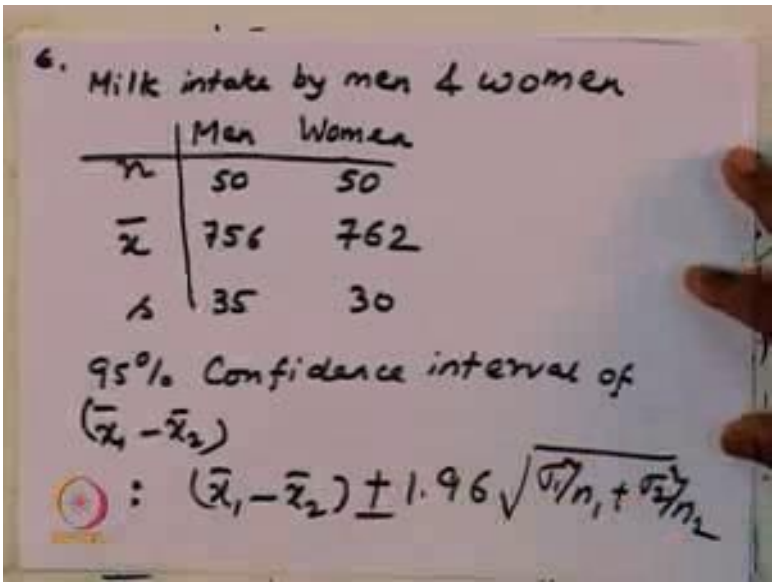
So, towards the end of last lecture we discussed about estimating the difference between 2 population means. We wanted to predict the difference between 2 population means. And what we said- if you have 2 distributions 2 populations which have means of mu 1 and variance of sigma 1 square, and another population with a mean of mu 2 and a variance of sigma 2 square. And if you draw a sample of size n 1, this is the population, and this is a sample size n 1 you get a mean of x 1 bar and a variance of s 1 square into x 2 bar s 2 square. So, the 95 percent confidence interval for; so x 1 bar minus x 2 bar is the best point estimate of mu 1 minus mu 2.

(Refer Slide Time: 07:43)



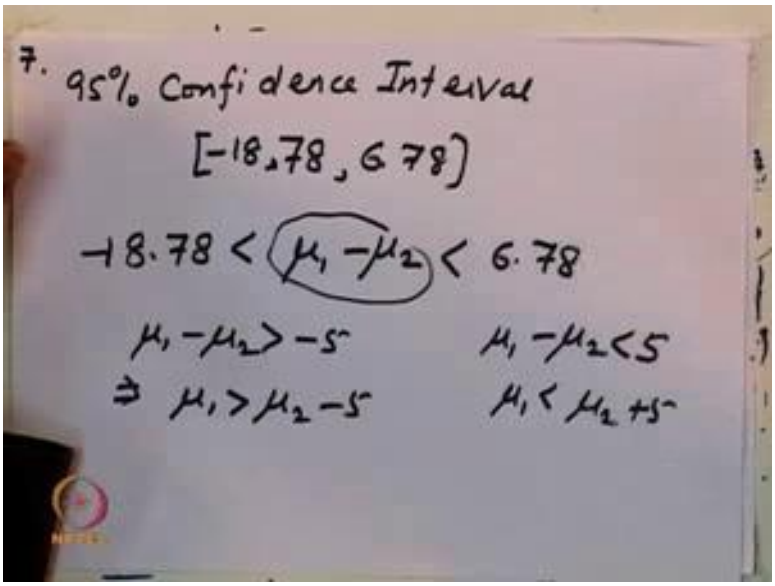
And I can write the variance. So, if I define x 1 minus x 2 as a random variable x (Refer Time: 07:51) random variable the variance will be given by; the variances of the individual terms which is sigma square by n 1 plus sigma 2 square by n 2. So, I can create this standard normal variable Z defined by x 1 bar minus x 2 bar minus of mu 1 minus mu 2 by root of sigma 1 square by n 1 plus sigma 2 square by n 2.

(Refer Slide Time: 08:41)



So, let us take a sample example. Imagine you want to know the milk composition; the milk intake by men and women. So, you have taken a sample. Let us say you chose 50 men and 50 women, and computed their average consumption of milk in a given week and let us say that is the average consumption. And the standard deviations are 35 and 30 respectively. You want to make a conclusion for determining whether men consume more men milk than women or the other way round. So, what we can do is we can find out the 95 percent confidence interval of x 1 bar minus x 2 bar and this confidence interval is given by x 1 bar minus x 2 bar plus minus. So, since you are provided standard deviation for the sample, so instead of sigma was you can use the sample standard deviations. And what you get out of it.

(Refer Slide Time: 10:34)

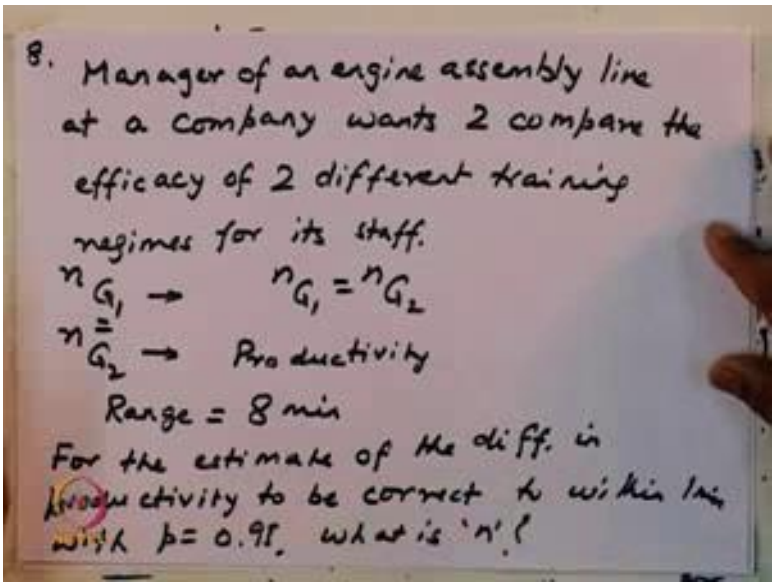


What you get out of it is this 95 percent confidence interval turns out to be minus 18.78 and 6.78. So, what this means is I can write minus 18.78 less than mu 1 minus mu 2 less than 6.78.

So, what it if mu 1 and equal to mu 2 was there? Then you would have concluded that there is no difference in the average intake of dairy products by men versus women. If it is negative it is greater than negative then it is a there is a possibility that your mu 1 minus mu 2 is greater than less a minus 5 implying mu 1 is greater than mu 2 minus 5. Or you could have had the other possibility mu minus mu 2 is less than 6. In that case mu 1 is less than mu 2 plus 5. So, there are both the possibilities where this entire difference is either negative or positive. So, this tells you it is very hard to conclude that the average intake of men and women of these dairy products really differ.

So, this is how you can calculate the confidence interval and draw some conclusions about the statistics. Let us solve one more example.

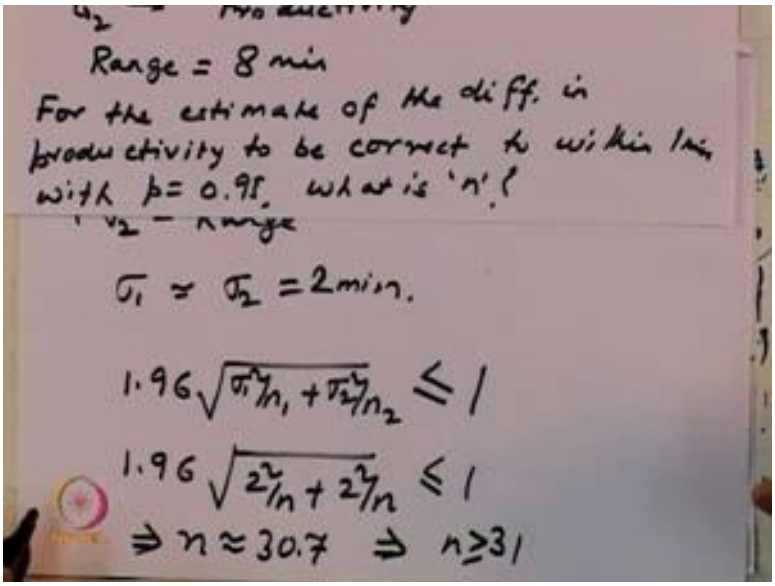
(Refer Slide Time: 12:12).



So, imagine the manager of an engine assembly line at a company wants to compare the efficacy of 2 different training regimes for its stuff. So, you choose two groups: G 1 and G 2 both of equal size. So, n of G 1 and n of G 2 are equal, you can write n of G 1 is equal to n of G 2. And after this you compare their productivity and the productivity what you are given is the difference so the estimate; so both these groups 8 minutes. So, if you are given the range of productivity for both these groups it is roughly 8 minutes, which means it takes between the minimum and the maximum time to perform a task it takes 8 minutes.

So what we want to know is, for the estimate of the difference in productivities to be correct to within 1 minute with p 0.95. What is n? In the other words what the question asked is how many people need to be in each group, so that when we compared the statistics of their performance, of the difference of that performance difference in the average performance the estimate of the difference in productivity is correct within 1 minute.

(Refer Slide Time: 15:07)

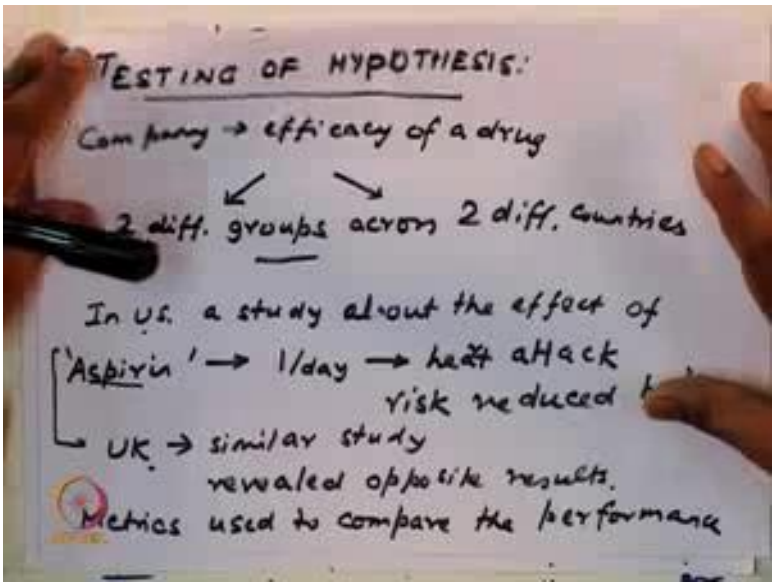


So what we are given, you have n of G 1 equal to n of G 2. What you are also given is the standard deviations, so sigma 1 is the stand is the standard deviation of the population for group 1 and sigma 2 is the standard deviation for population of G 2. Now, you remember this approximation as 4 times sigma is equal to range. So, I can write 4 times sigma equal to range, and this I am given as 8 minutes. So, from this you can estimate sigma 1 equal to sigma 2 equal to 2 minutes.

And you want the difference in productivity to be correct within 1 minute: so you want the difference in productivity to be correct within 1 minute. So, the difference in productivity is given by 1.96 times and this is less or equal to 1 minute. In the above expression you can simplify it by writing 1.96 root of 2 square by n plus 2 square by n assuming n G 1 equal to m G 2 less equal to 1. And this gives, if you solve this equation you get n roughly equal to 30.7. So, this means your n has to be greater than 31 greater or equal to 31.

So, this concludes our discussion on confidence intervals, we will next discuss an important concept where we make use of these confidence intervals or inputs from this and that is the statistical testing of hypothesis; this is called Testing of Hypothesis.

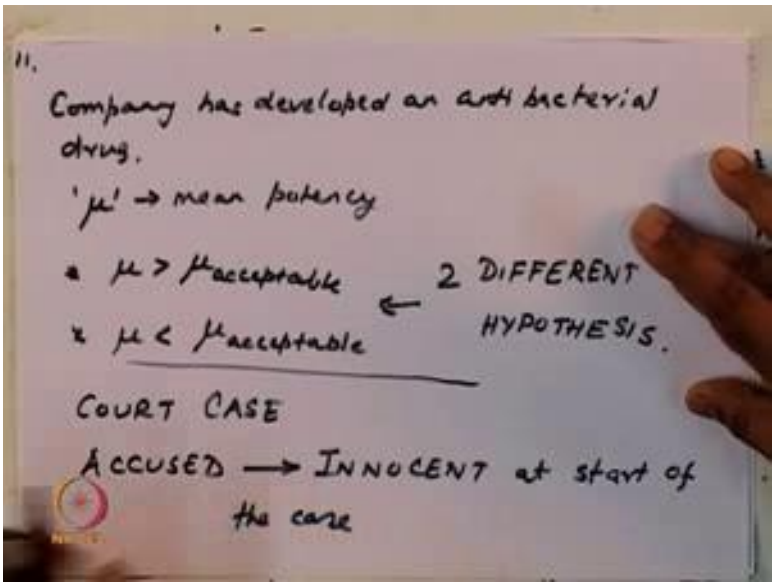
(Refer Slide Time: 17:16)



Imagine, when there is a pharmaceutical company is deciding on developing a drug. You have a company which tests the efficacy of a drug and imagine that this particular efficacy is tested in two confuse or to 2 different scenarios, they are by 2 different groups across. So, ideally the efficacy of the drug should be identical irrespective of which group is do it, which means the conditions it sets to compare the performance of the drug has to be the same. But what if it is different? That would be very difficult, because the very idea or knowledge of whether the drug works or not is not clear. So, this is where the testing of hypothesis is very important.

As an example, in US a study concluded a study about the effect of the drug call aspirin. This drug concluded that if a patient was taking 1 pill per day, this leads to reduction of heart risk or a heart attack risk reduced by half. This is a very strong finding that this really says that this drug has an amazing effect in reducing the risk of heart attack. Neutrally, a very short period after the study was come came out in UK a similar study revealed opposite results. So this is very intriguing; how is it possible that the same drug is working in one country and not in the other country? It is debatable whether this is because of differences of people, but more likely it is because of the differences in the matrix used; so the matrix used to compare the performance. So, this is where statistics becomes very important.

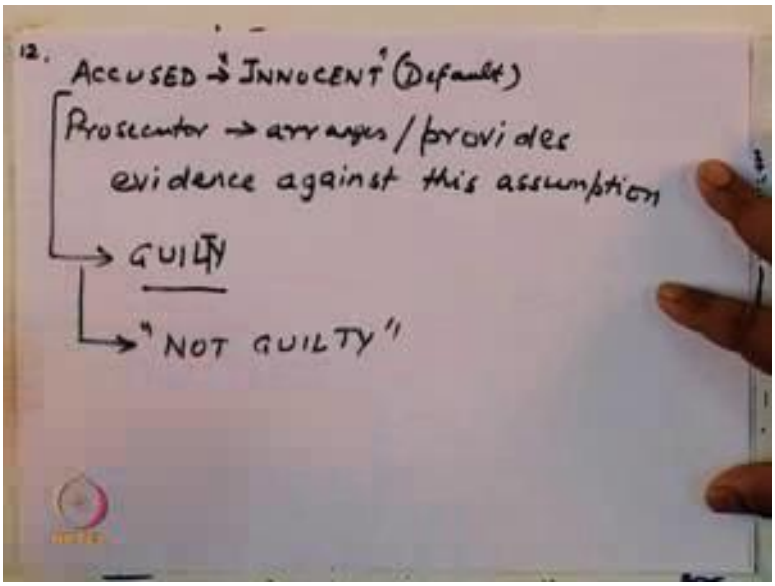
(Refer Slide Time: 20:39)



And in testing of hypothesis what is typically done. So, let us say you know again let us take another example where a pharmaceutical company again; a company has developed an antibacterial drug. So, for this drug the company can go ahead and find out what is just mean potency, this is the mean potency of the drug. And the company needs to know whether this mean potency is acceptable as per government norms. So, instead of actually calculating its means potency what the company can actually do is whether it can raise or address two possibilities. Possibility number 1 saying- the average potency mu is greater than mu acceptable; whatever government norms stipulate that this potency in is beyond this. Or the other possibility is mu is less than mu acceptable.

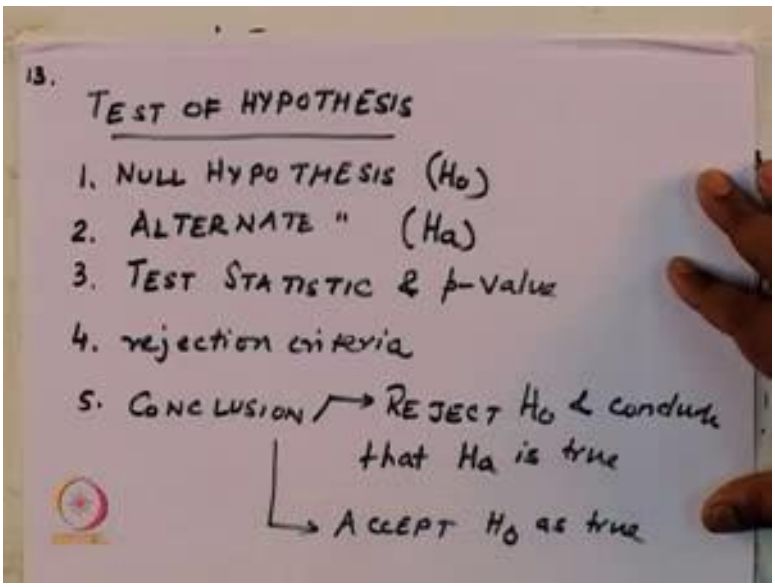
So these are two different hypotheses. So, the way to go about testing this hypothesis is very similar to a court case. You have a person accused who by default is assumed to be innocent. So, the accused is always assumed to be innocent at start of the case.

(Refer Slide Time: 22:55)



So you have the accused who is assumed innocent, and this is always the default. And what the prosecutor does? Arranges evidence, provides evidence against this assumption. So, if the evidence is significant then the accused is pronounced guilty by the lawyer. But if the evidence is not substantial, then the person is declared not guilty. So, note the difference; not guilty and innocent is not the same. Just says that there was not sufficient evidence to label this person as guilty. So, how do we go about doing a testing of hypothesis?

(Refer Slide Time: 24:29)



A test of hypothesis has five components: one is called the null hypothesis; just typically represented by H naught, H naught represents the null hypothesis. You have an alternate hypothesis, this is typically represented as H a. So, based on the evidence you calculate something which is called a test statistic, and based on that using the test statistics you calculate something called a p value. You decide on some reaction or acceptance criteria reaction criteria. And 5, you draw your conclusion. So there are two possibilities: you can either reject H naught or conclude that H a is true or you can accept H naught as true.

With that I would complete my lecture today. And in the next class we will continue where we left off. We will take up some examples of test of hypothesis and see how we go about defining the test statistics, how we go about defining the null hypothesis, alternate hypothesis and so on and so forth.

Thank you for your attention.